Improving ML training sets with Active Learning in FINK

ML4Astro Catania 2024

Anais Möller
ARC DECRA Fellow

Australian Government
Australian Research Council

Acknowledging the traditional owners of Swinburne's land, Wurundjeri People of the Kulin Nation

SWINBURNE UNIVERSITY OF TECHNOLOGY

Rubin LSST

Now

Raw Data
Sequential 30s image, 20TB/night

60s

Prompt Data Product
Difference Image Analysis
Alerts: up to 10 million per night

Public data!

24h

Prompt Products DataBase
Images, Object and Source catalogs from DIA
Orbit catalog for ~6 million Solar System bodies

Year

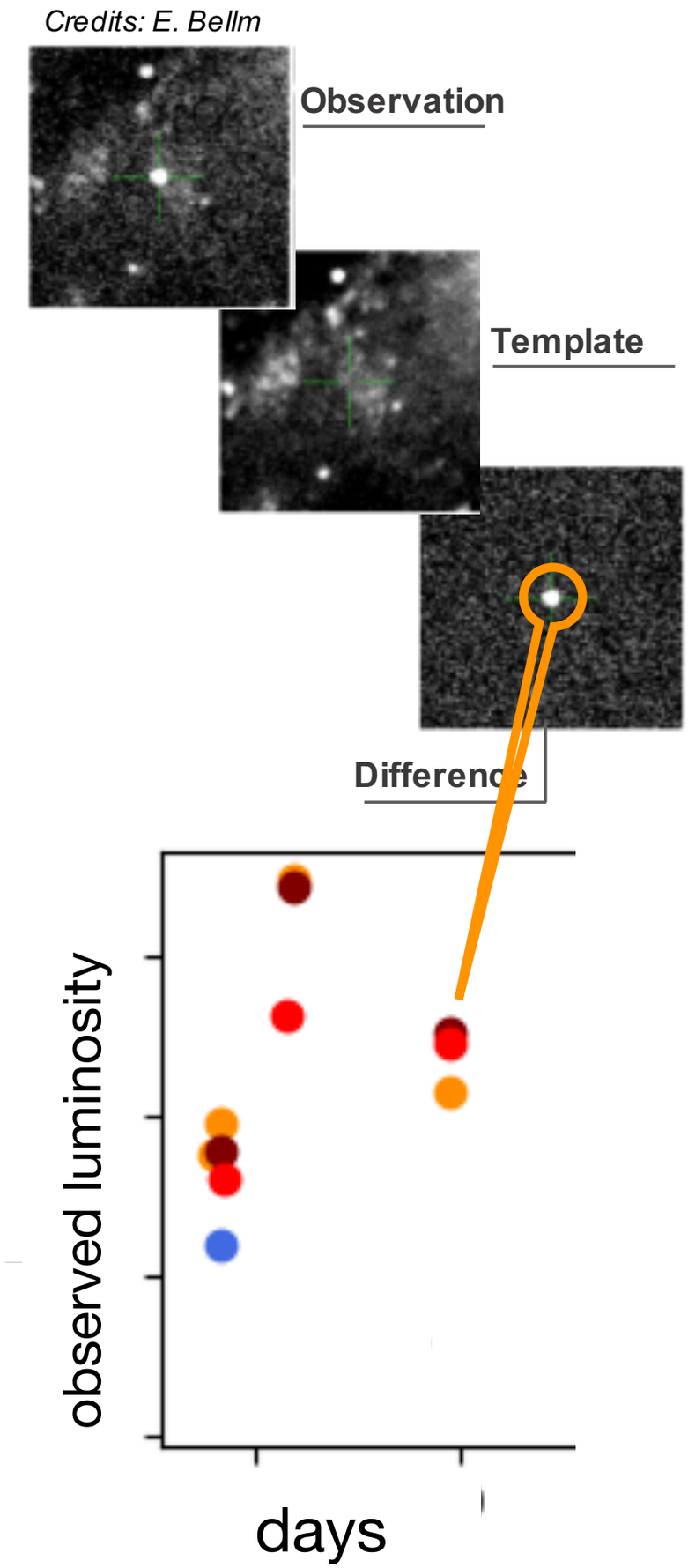
Annual Data Release
Accessible via the LSST Science Platform &
LSST Data Access Centers.

End

Final 10yr Data Release
Images: 5.5 million x 3.2 Gpx
Catalog: 15PB, 37 billion objects

LSST PIs and teams

Credits: E. Bellm



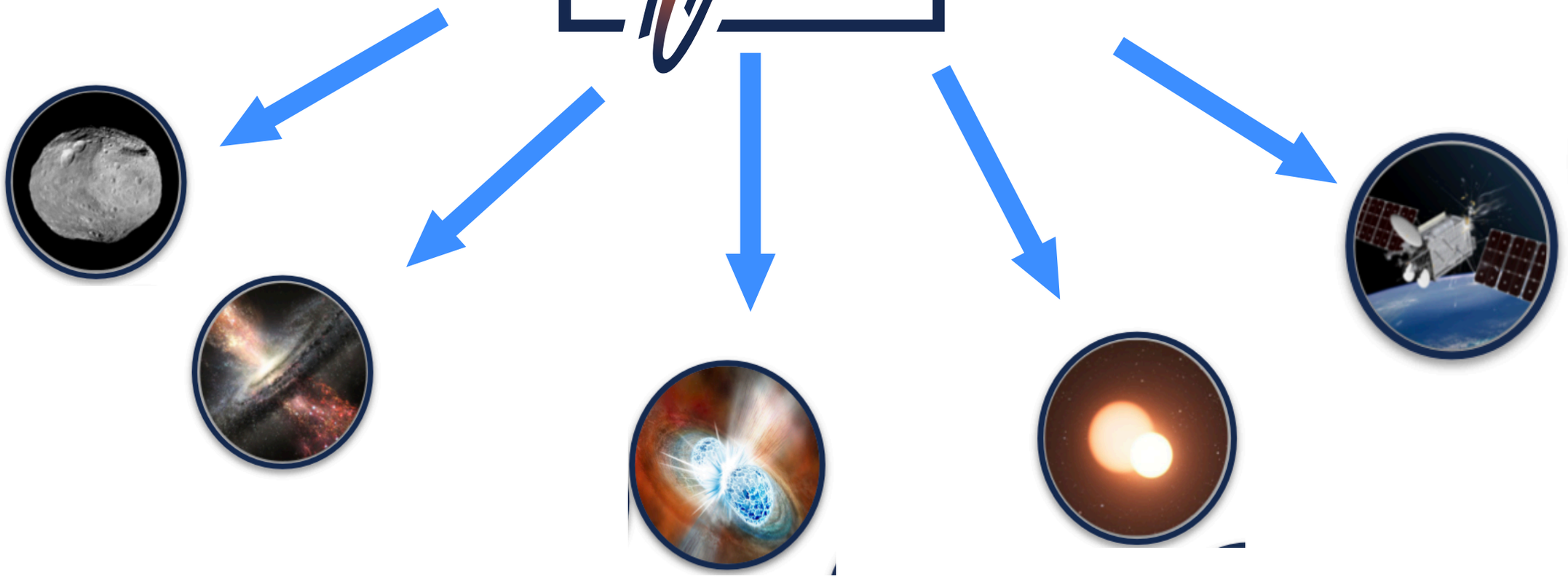
Rubin LSST

10 million t-domain candidates/night

Brokers such as

FINK

Hosted at CC-IN2P3 for Rubin



Möller, Peloton, Ishida +2020

LAPP, Annecy - 19 and 20 May 2022

2nd Fink Collaboration Meeting

8 - 10 January 2024, IJCLab - France

Photo: Carl Knox – OzGrav, Centre of Excellence for Gravitational Wave Discovery

3-5 May 2023, Melbourne - Australia

6-10 May 2024 - CBPF, Rio de Janeiro - Brazil

Fink-Brazil Workshop

broad-science collaboration

community driven, open to everyone

PIs & management team Peloton, Ishida, AM

FINK

Community driven, join us!

Supernovae

SNIa, SN, PISN

Allam+2023, Leoni+2022, Möller+2022
CNRS MITI grant

Kilonovae

Orphan, GW+KN, Fast transients

Grandma+Fink 2022,2023
Biswas+2022,2023

Nuclear transients

AGN, TDE

Russeil+2022
Gondhalekar+ in prep.

Gamma Ray Bursts

Solar System Objects

Discovery, tracking

Le Montagner+ 2023
Carry+2024
CNRS MITI grant

Space Awareness

Satellite glints

Karpov+2023,2022

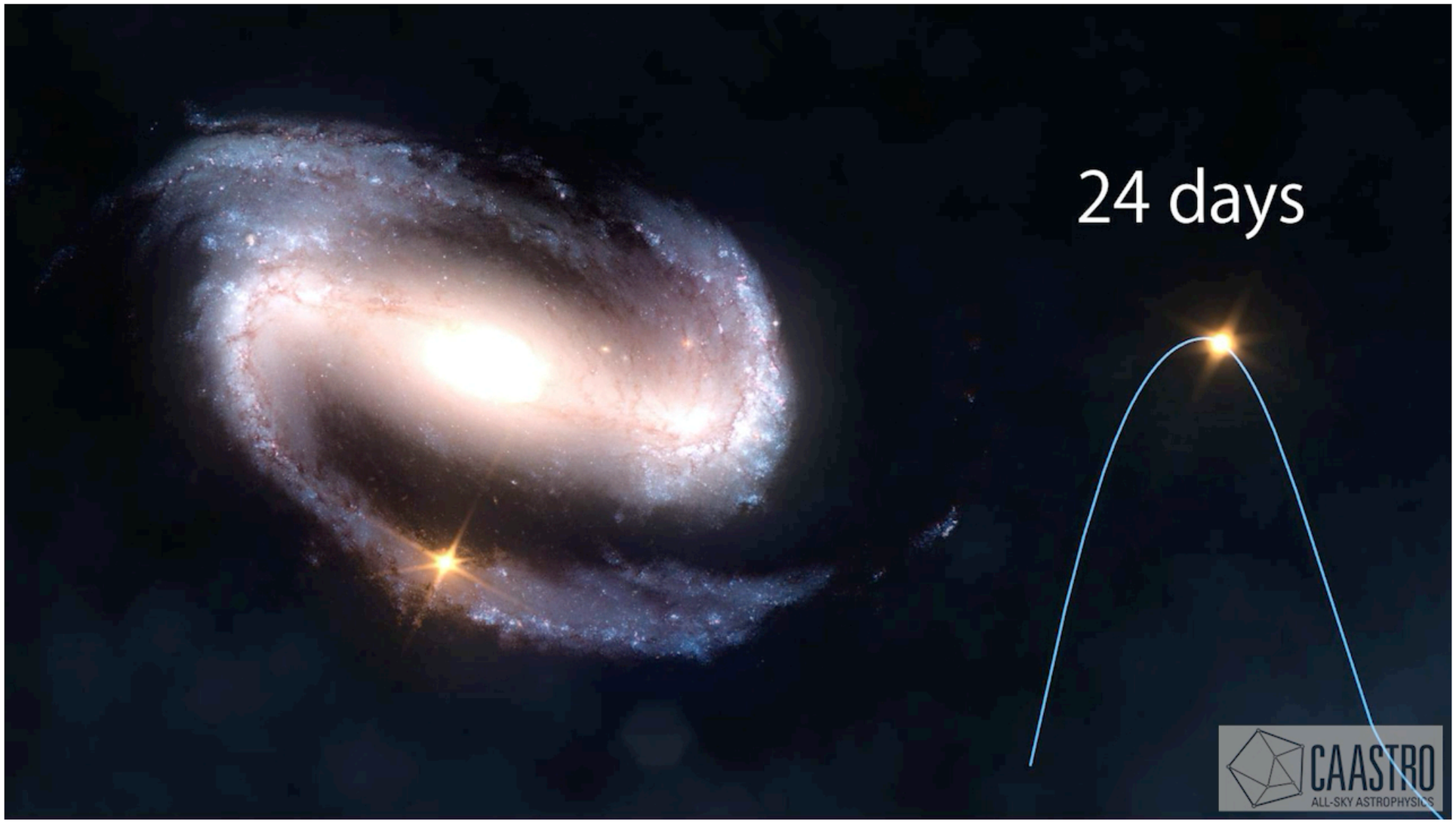
Anomalies

Pruzhinskaya+ in prep.

LSST simulations: ELAsTiCC

Fraga+ 2024

Type Ia supernovae

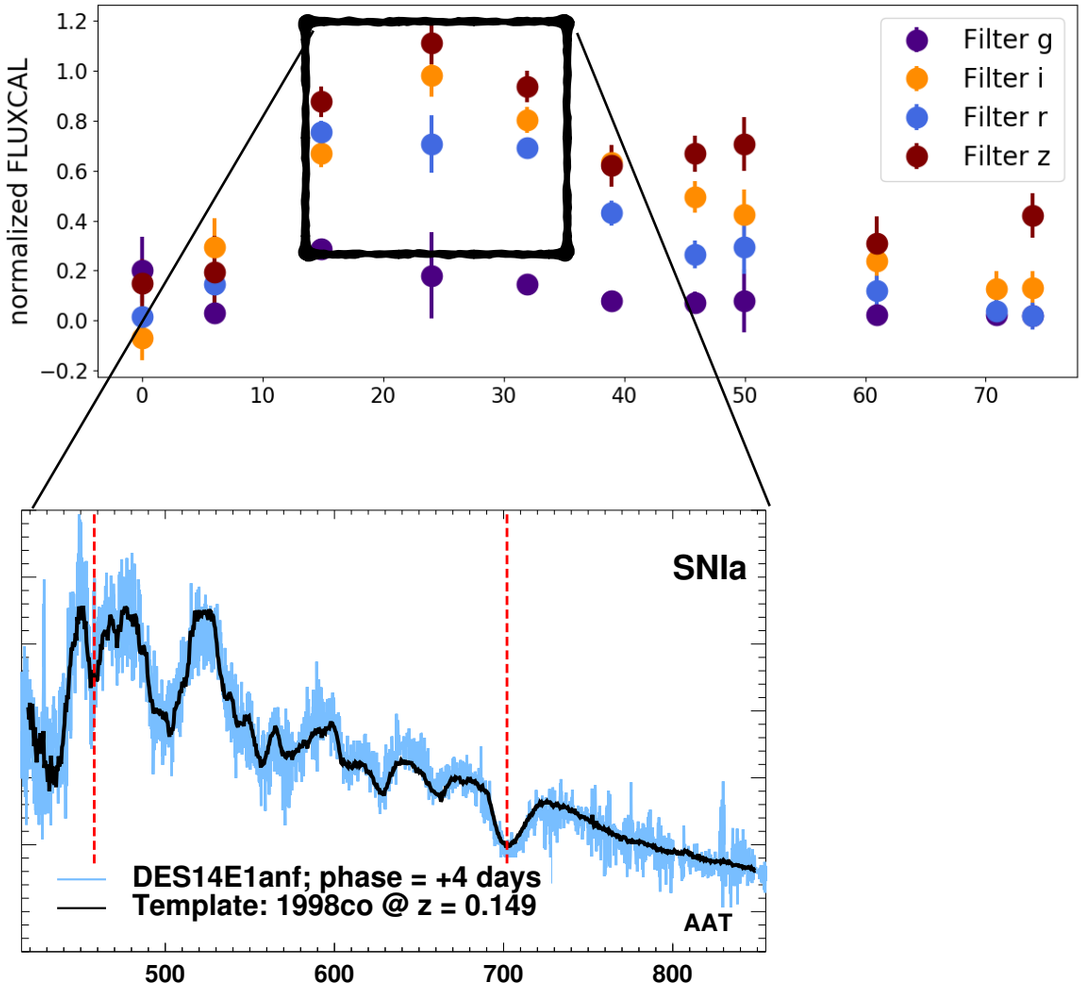


Homogeneous spectral and photometric properties

Rubin > 1 million

Direct measurement cosmic expansion

DES ~ 2k

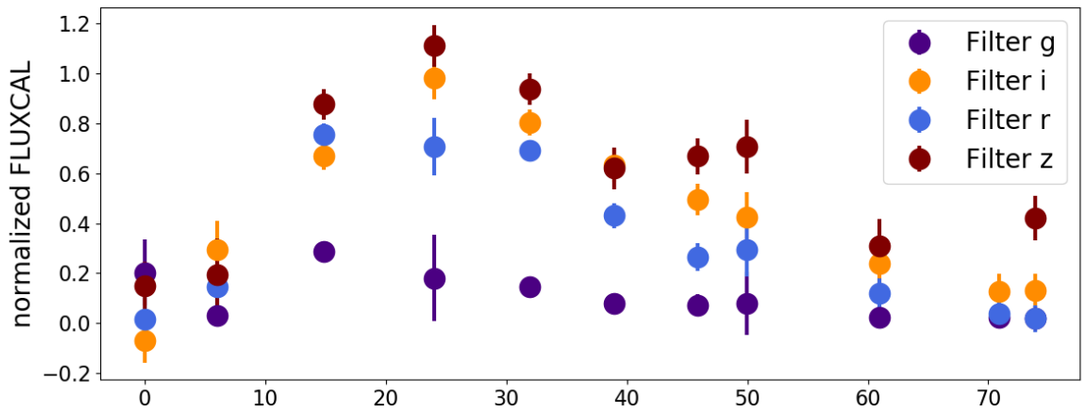


DARK ENERGY SURVEY

415 SNe Ia

(expected 2,000 detected)

Spectroscopic classification



ML algorithm

Photometric classification

1. Supervised + simulations

2. Unsupervised

3. AL for improving training sets

1. **Supervised + simulations**
2. Unsupervised
3. AL for improving training sets

SuperNNova

Recurrent Neural Networks

Bayesian Neural Networks

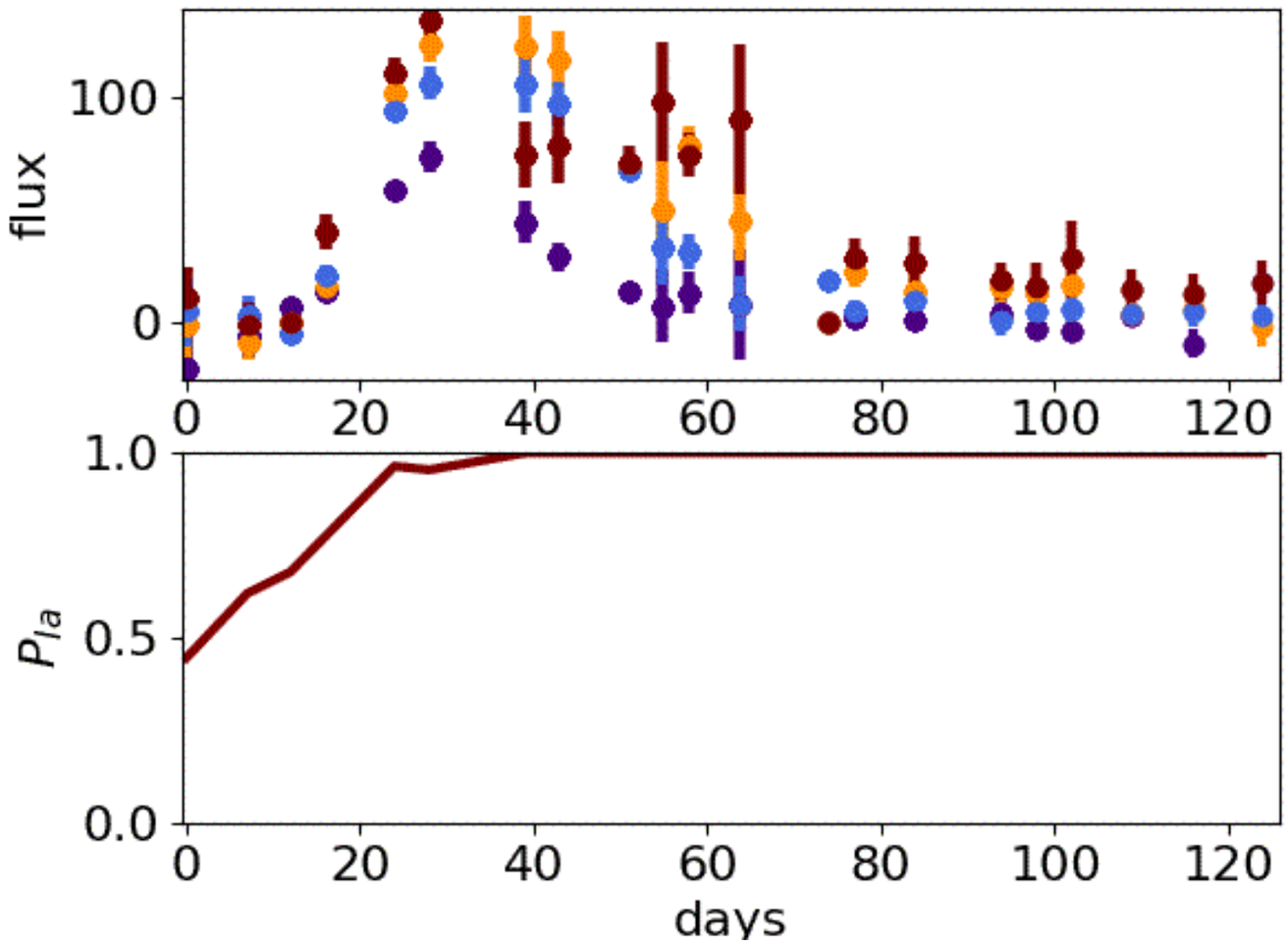
(MC dropout & Bayes by Backdrop)

Möller+ 2019, 2022b

Handles irregular time series

Used with DES, PS, ZTF data

Soon Rubin!

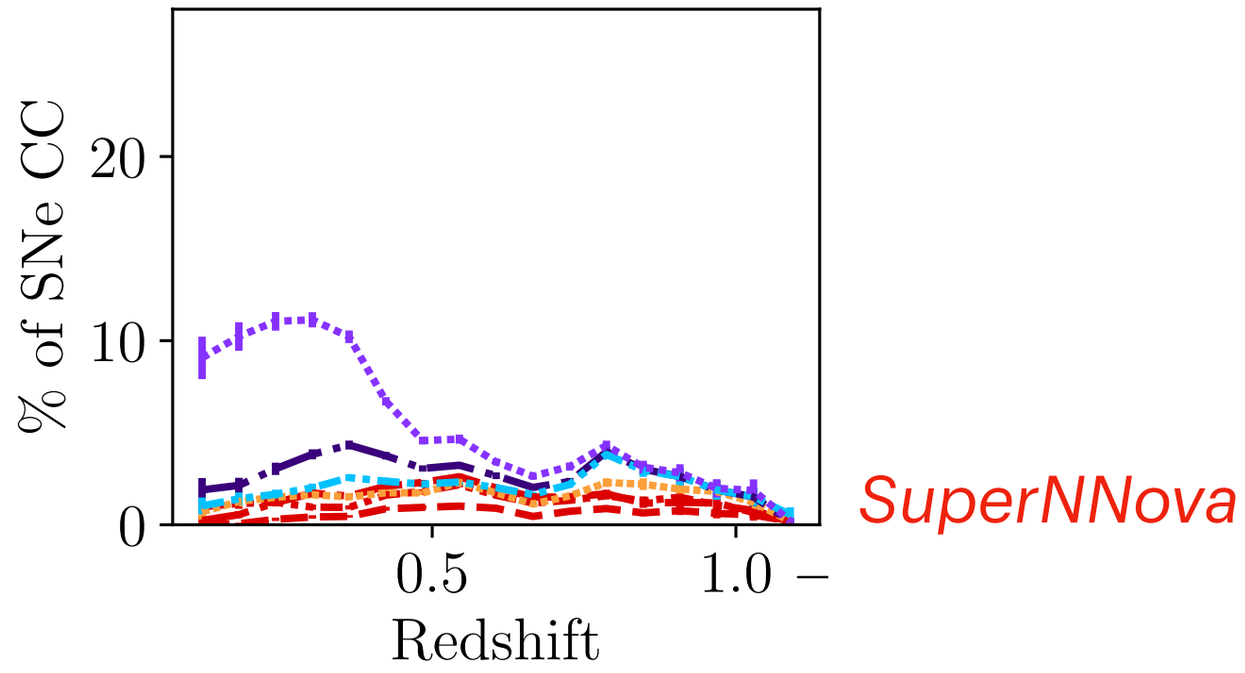


SuperNNova

Möller+2019

Pytorch, many configurations possible

DES: Light-curves + host-galaxy redshifts

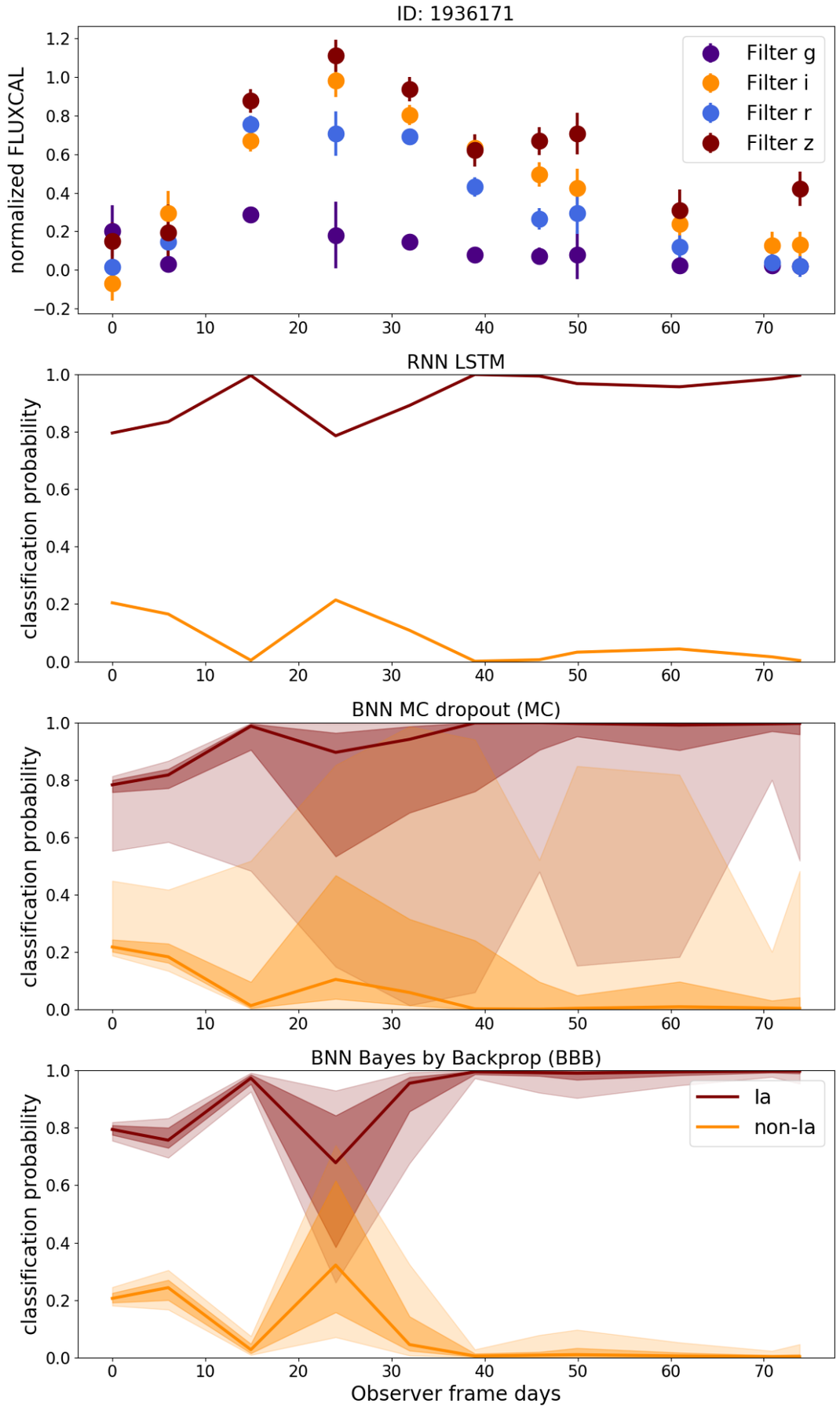


Vincenzi, Sullivan, Möller et al. 2021

State-of-the-art simulations

Vincenzi+2020

>98% accuracy


ID: 1936171
normalized FLUXCAL
Filter g
Filter i
Filter r
Filter z
RNN LSTM
classification probability
BNN MC dropout (MC)
BNN Bayes by Backprop (BBB)
Ia
non-Ia
Observer frame days

Möller+2022

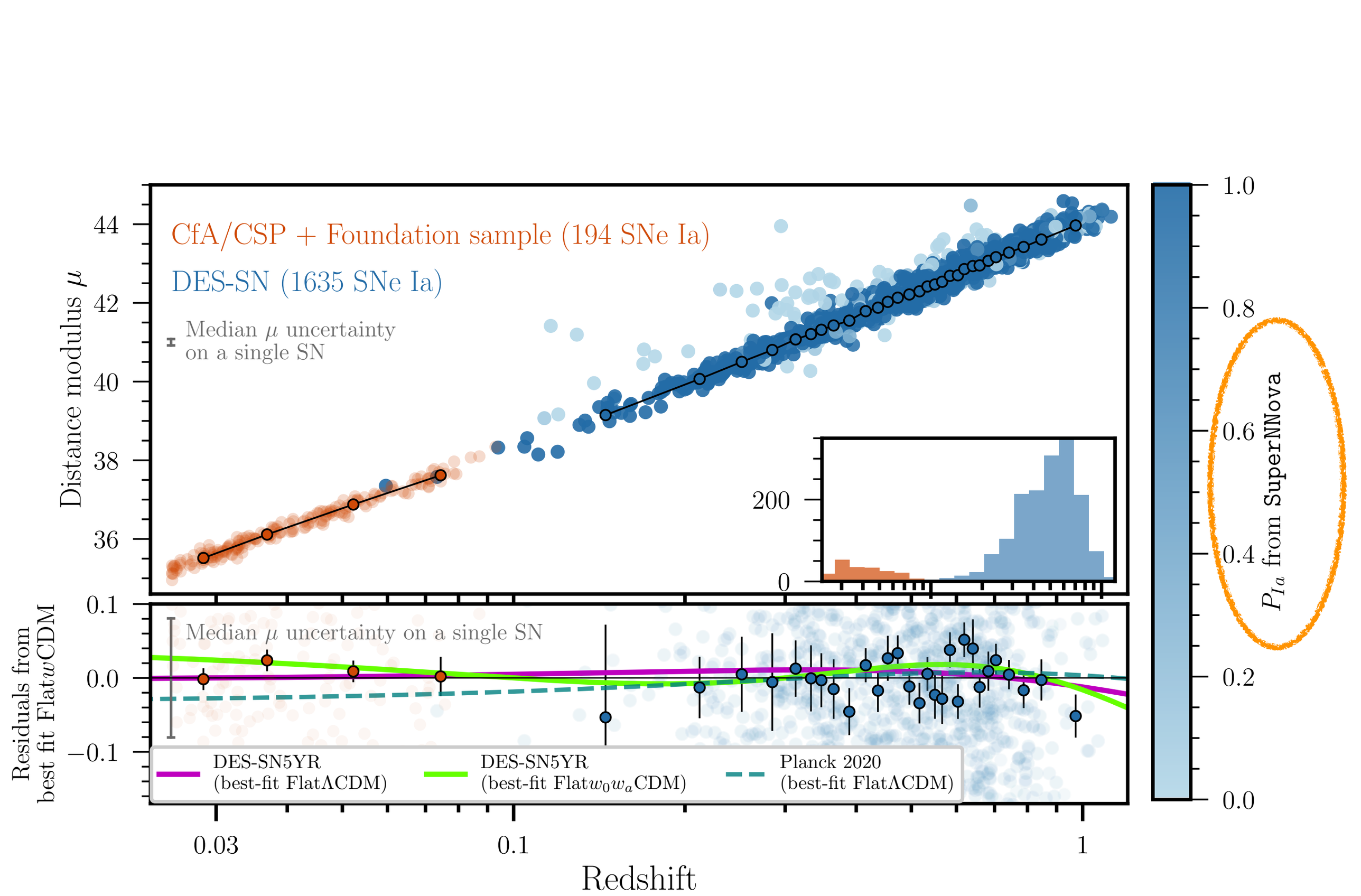
DES: Light-curves + host-galaxy redshifts

1484 SNe Ia

Largest high-z SN Ia sample from a single survey for cosmology

Möller+ 2022a, Vincenzi+ 2024, DES+ 2024

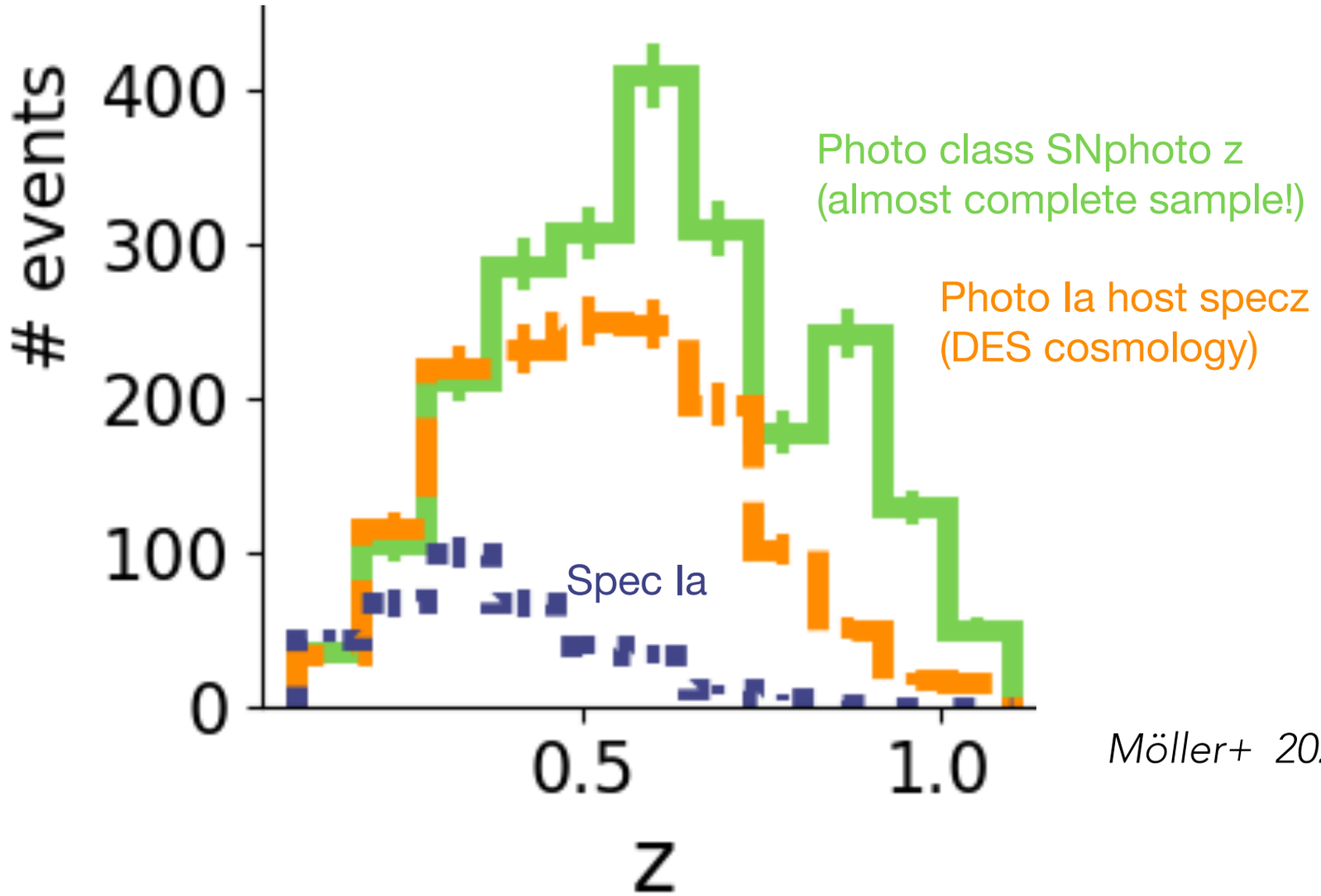
DES Collaboration 2024



DES Collaboration 2024

DES: Light-curves + ~~host-galaxy redshifts~~

DES: Light-curves + ~~host-galaxy redshifts~~



Möller+ 2024

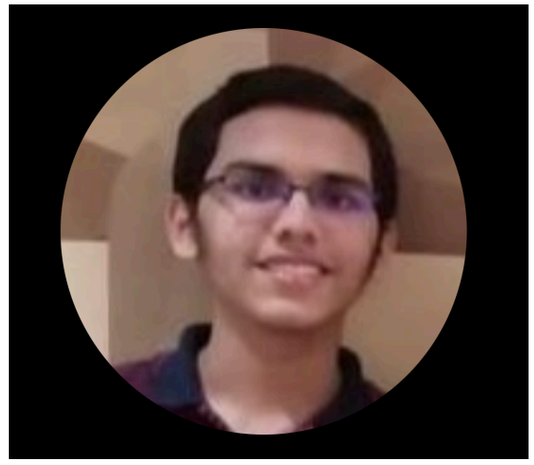
2,411 high quality SNe Ia

1. Supervised + simulations

2. Unsupervised

3. AL for improving training sets

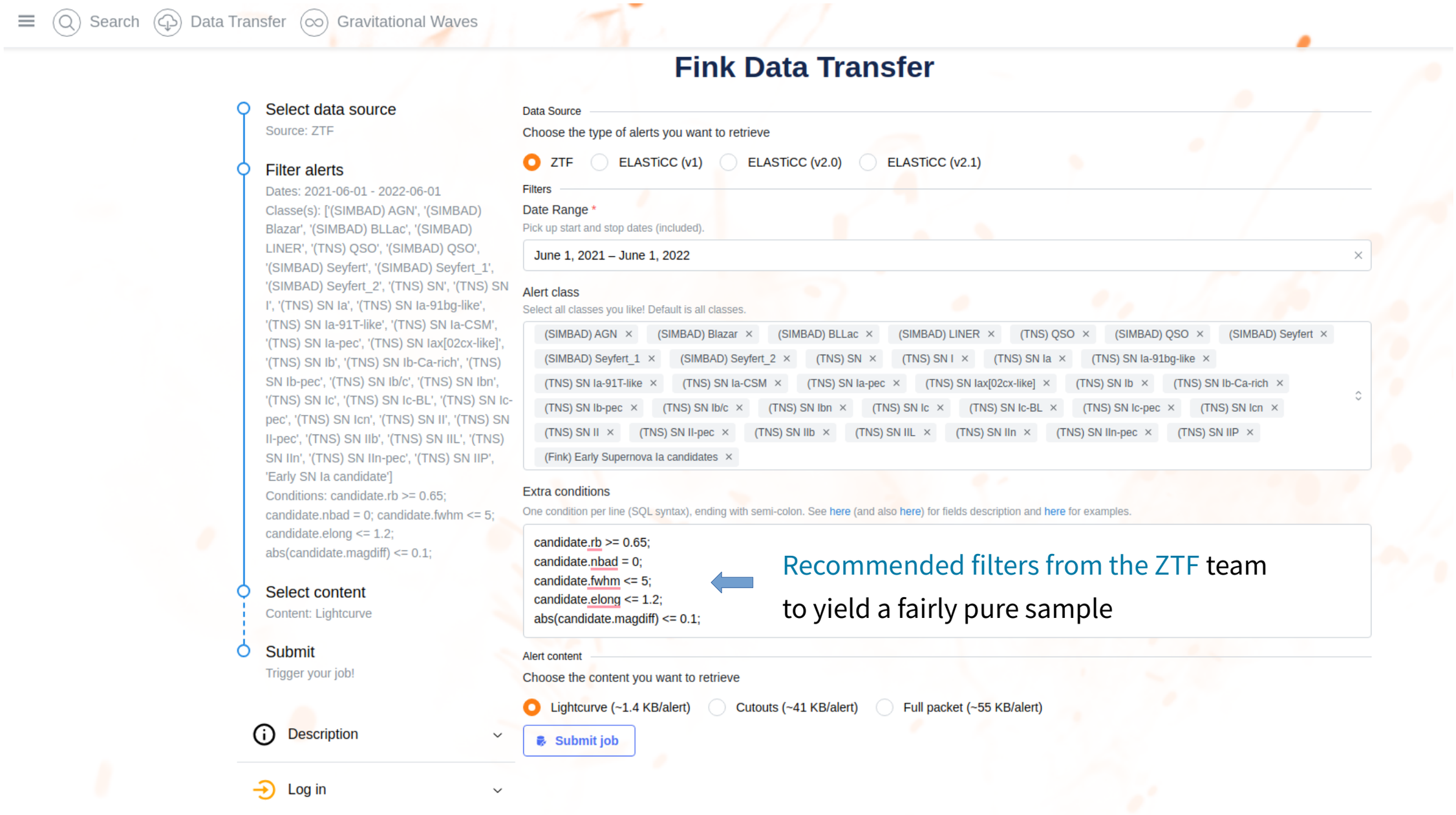
Unsupervised SN vs AGNs



Preliminary!!!

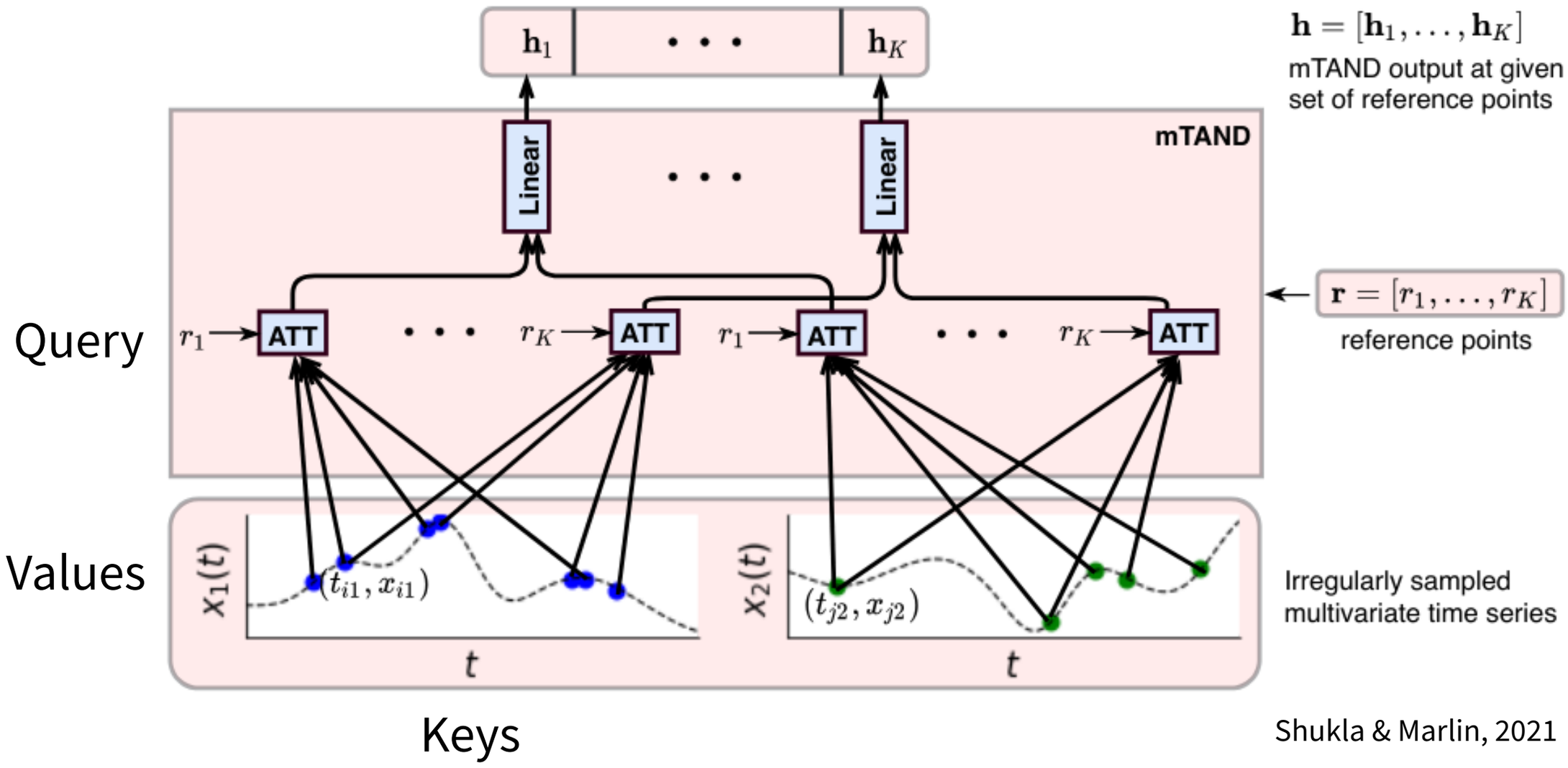
Gondhalekar+ in prep.

https://fink-portal.org/download



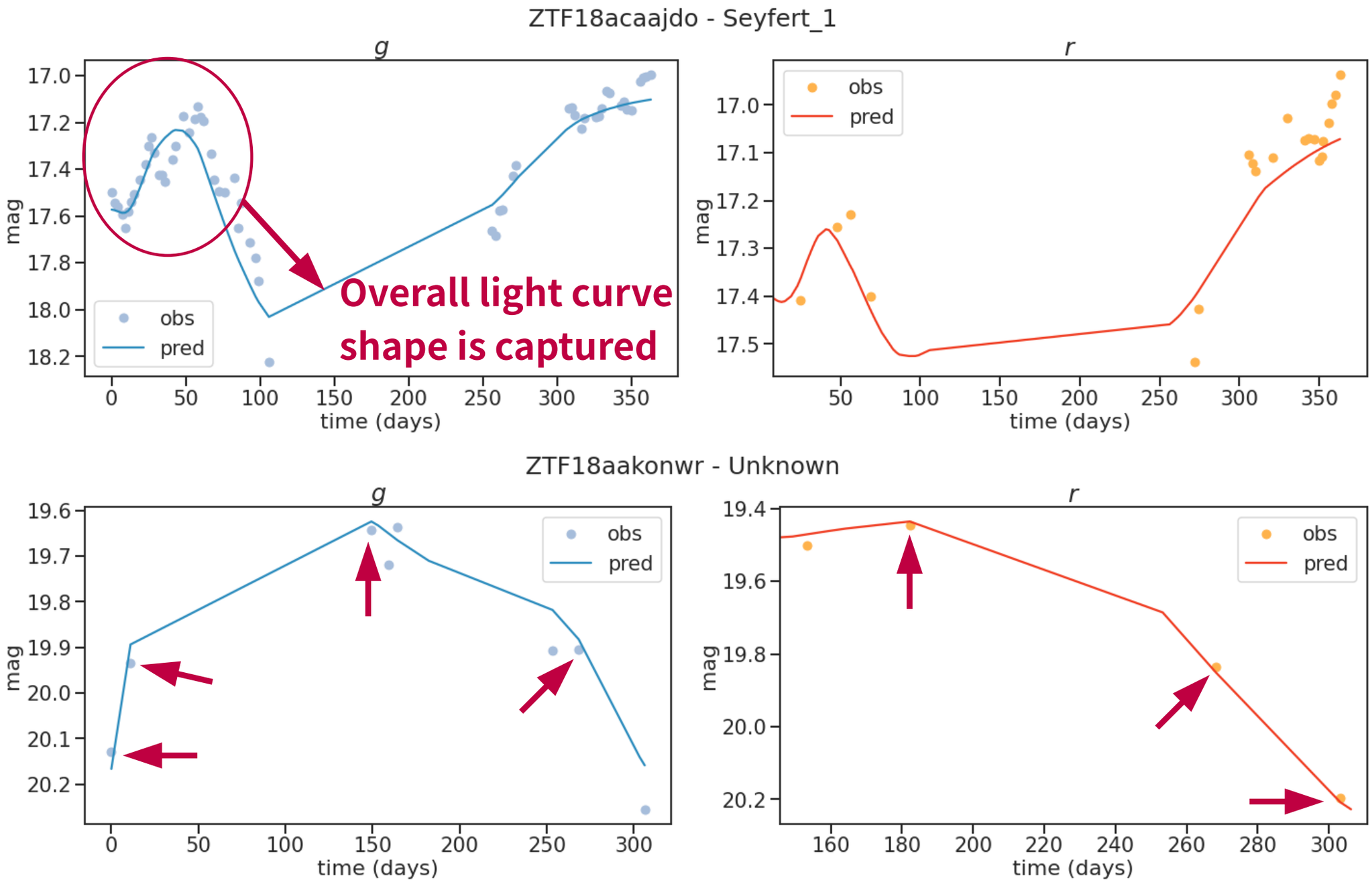
Gondhalekar+ in prep.

mTAN: Multi-Time Attention Network



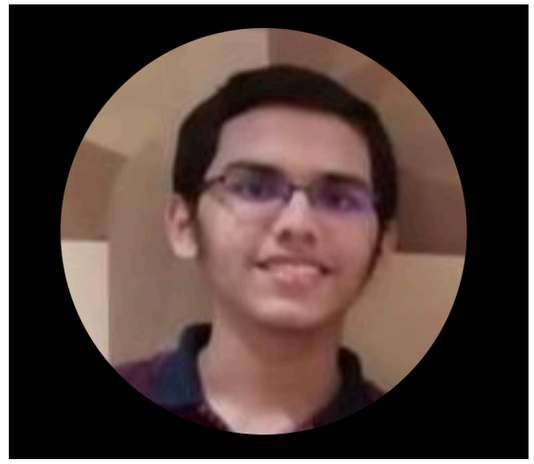
Shukla & Marlin, 2021

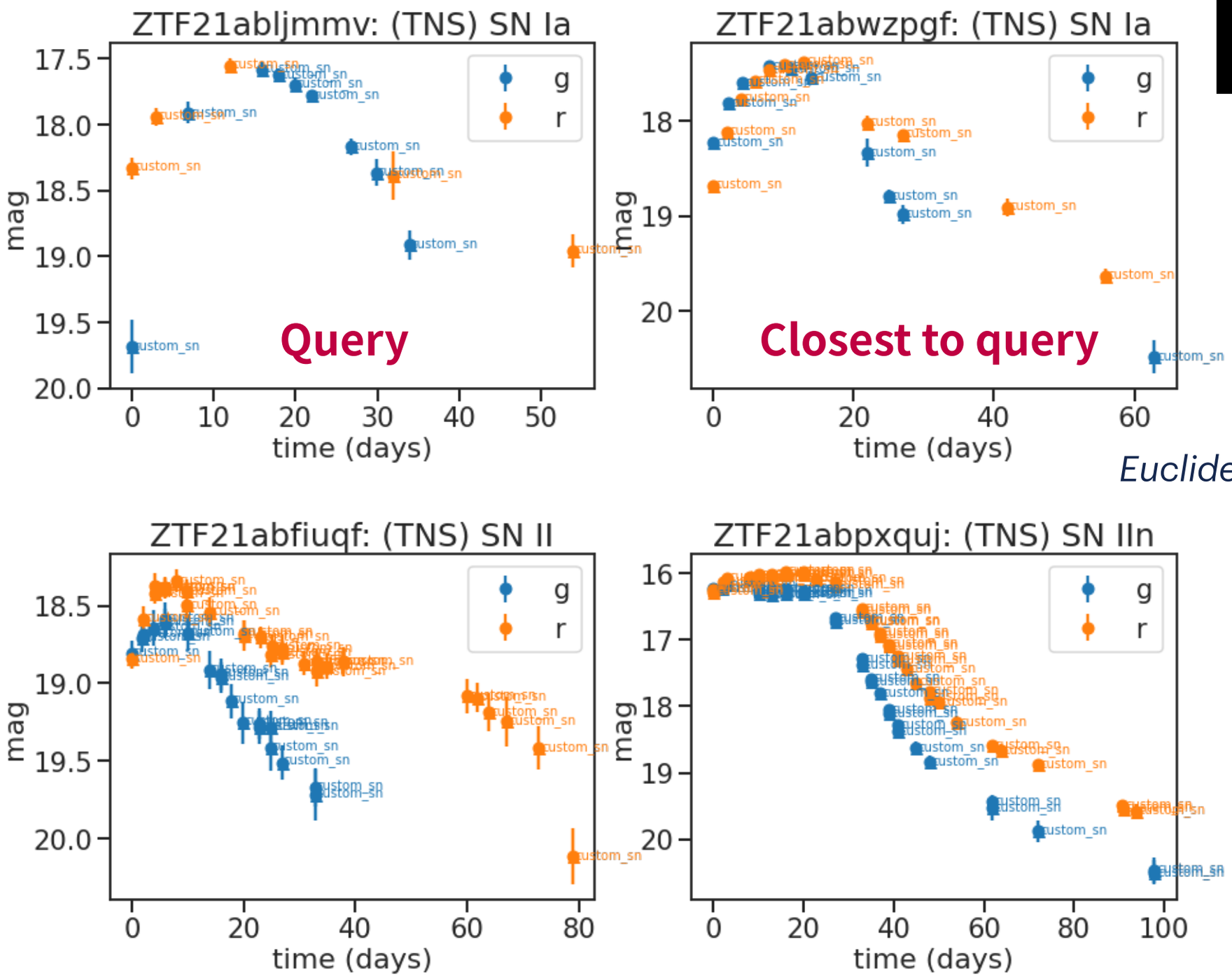
Preliminary!!!



Gondhalekar+ in prep.

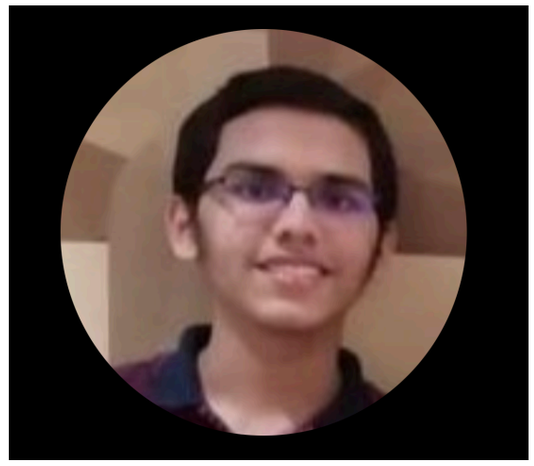
Preliminary!!!





Euclidean, cosine distance....

Gondhalekar+ in prep.



~0.019 s /light-curve

Compared to the time-series transformer model of Allam, Jr. & McEwen (2021), this is*:

1. ~80 times faster inference (0.019s vs. 1.5s).
2. ~3 times smaller model size (340 KB vs. 1100 KB).

Assuming ~1 million light curves/year, this amounts to ~0.34 kg of CO_2, equivalent to driving 1.37 km by an average ICE car**.

*The follow-up paper (Allam, Jr. et al 2023) reduced model size by ~18 and execution time by ~8 times than their original approach.

**Source: MachineLearning Impact calculator

Gondhalekar+ in prep.

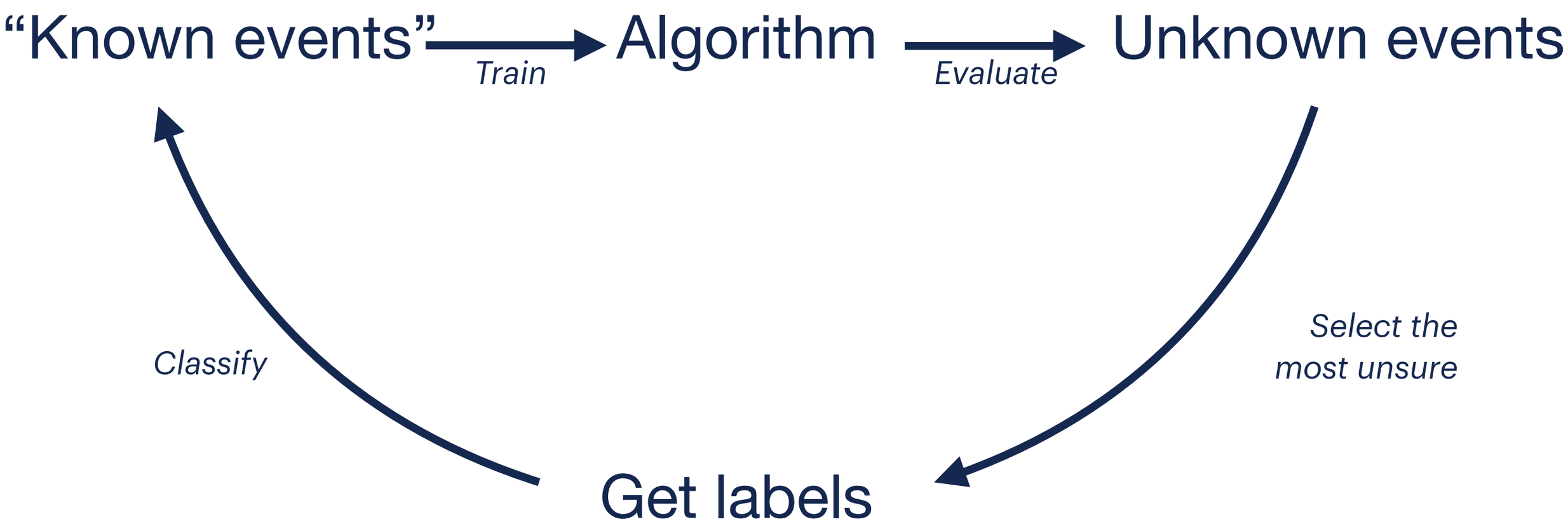
1. Supervised + simulations

2. Unsupervised

3. AL for improving training sets


“Known events”
Train
Algorithm
Evaluate
Unknown events
Select the best
Your target
(e.g. SNIa)

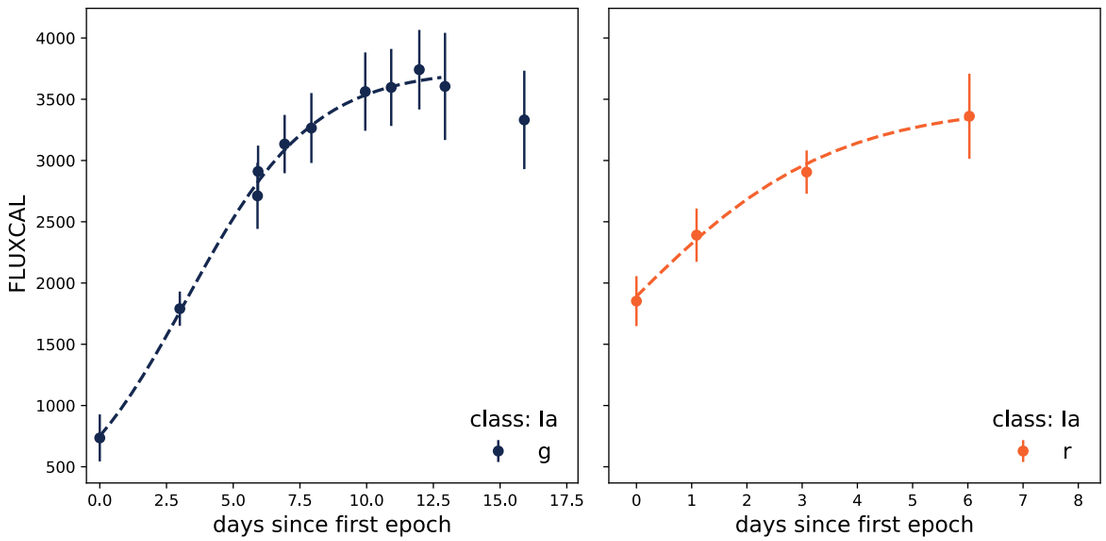
Active Learning



Early SNe Ia

We don’t know SNe Ia progenitors

Missing modelling for early SNe Ia



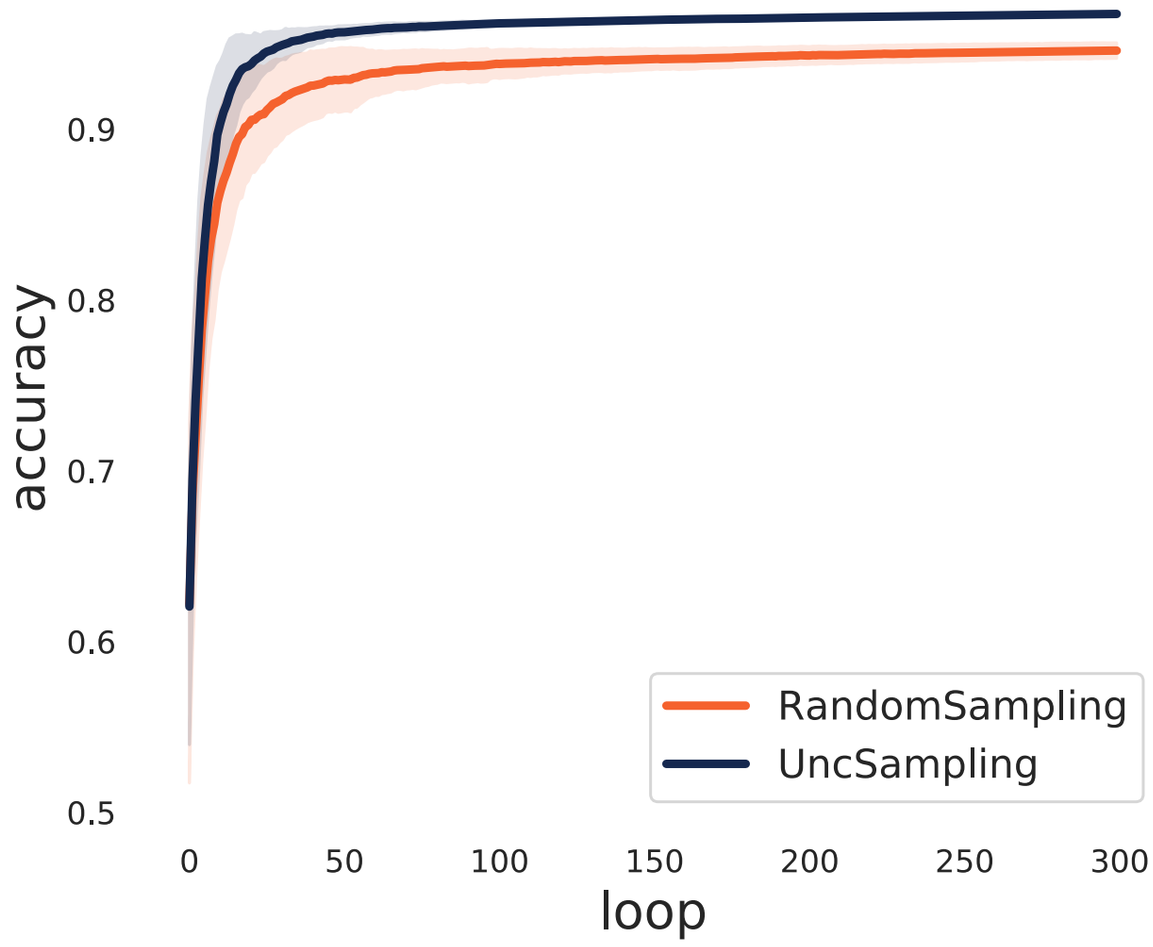
Features: sigmoid fit

Random Forest

Ishida+2019

Leoni, Ishida+ 2022

Active Learning SN classification (simulated)



Leoni, Ishida et al. 2022

Active Learning SN classification (real data)

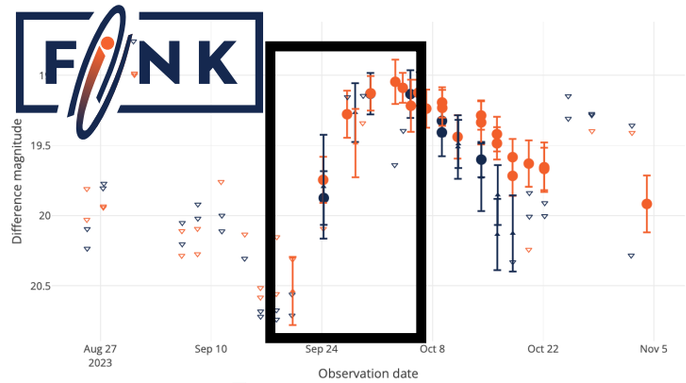
~30 known lcs

Leoni, Ishida et al. 2022

Train

RF

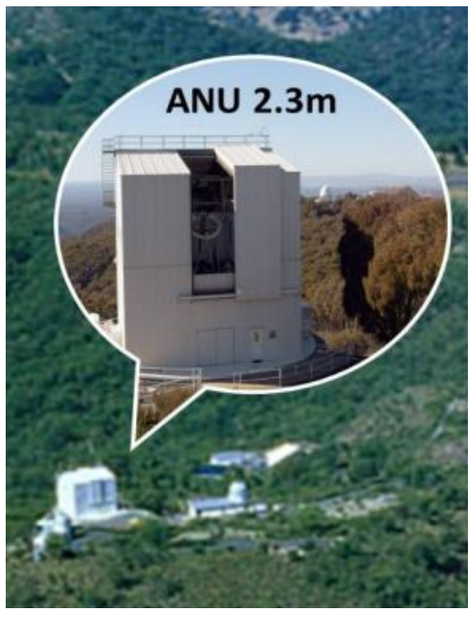
Evaluate



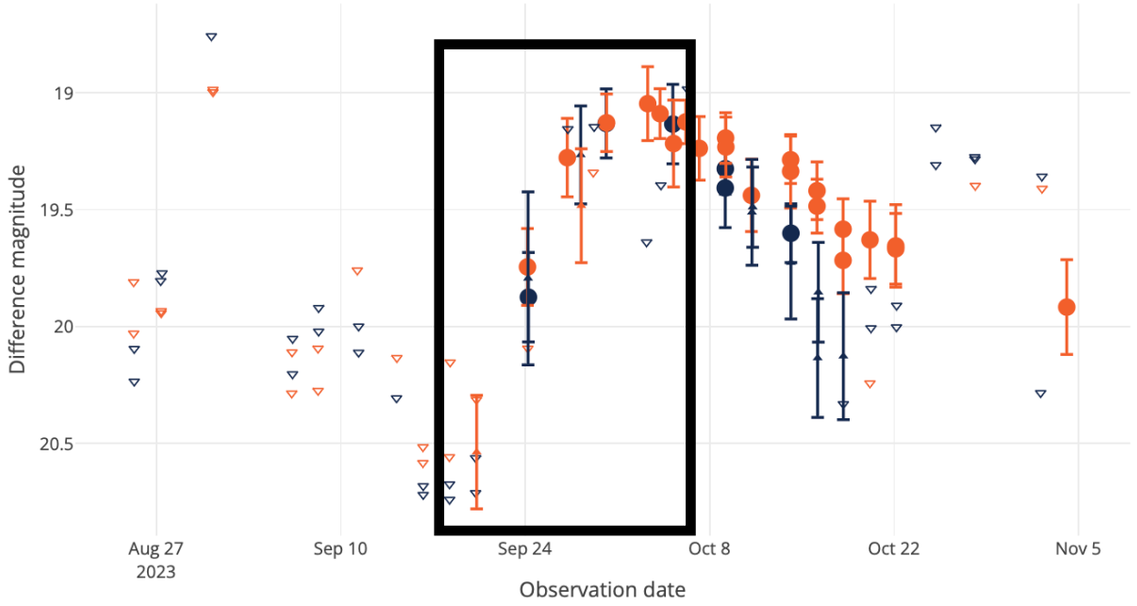
Select the most unsure

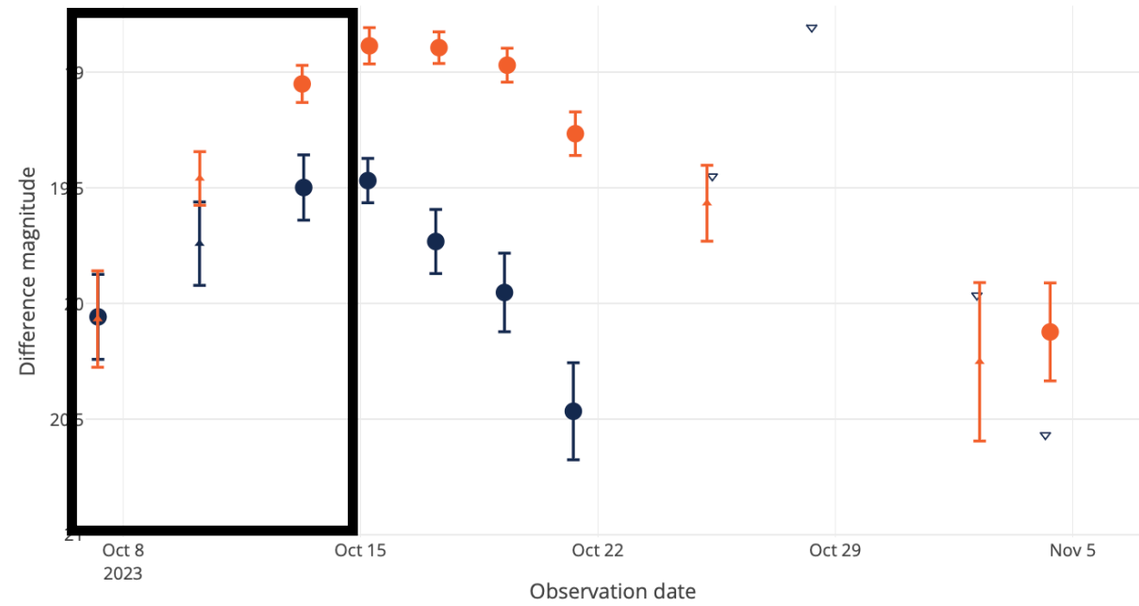
closest to P_{Ia} = 0.5

Get classification



Spectroscopy

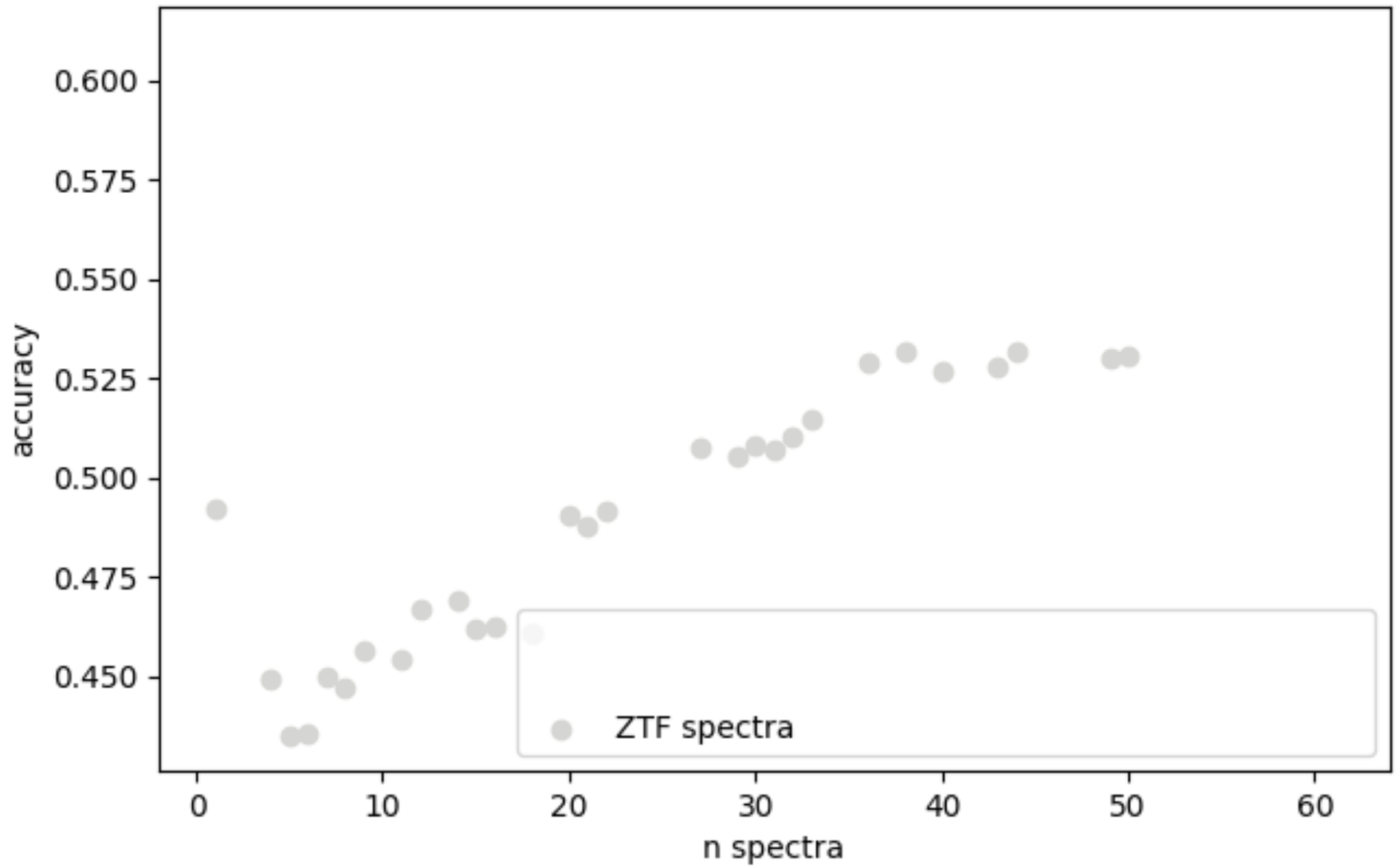




2 CV, 16 SN Ia, 5 SN II,1 SN Ib, 2 SN Ibn, 1 SN Ic, 1 microlensing event…

Möller et al. in prep.

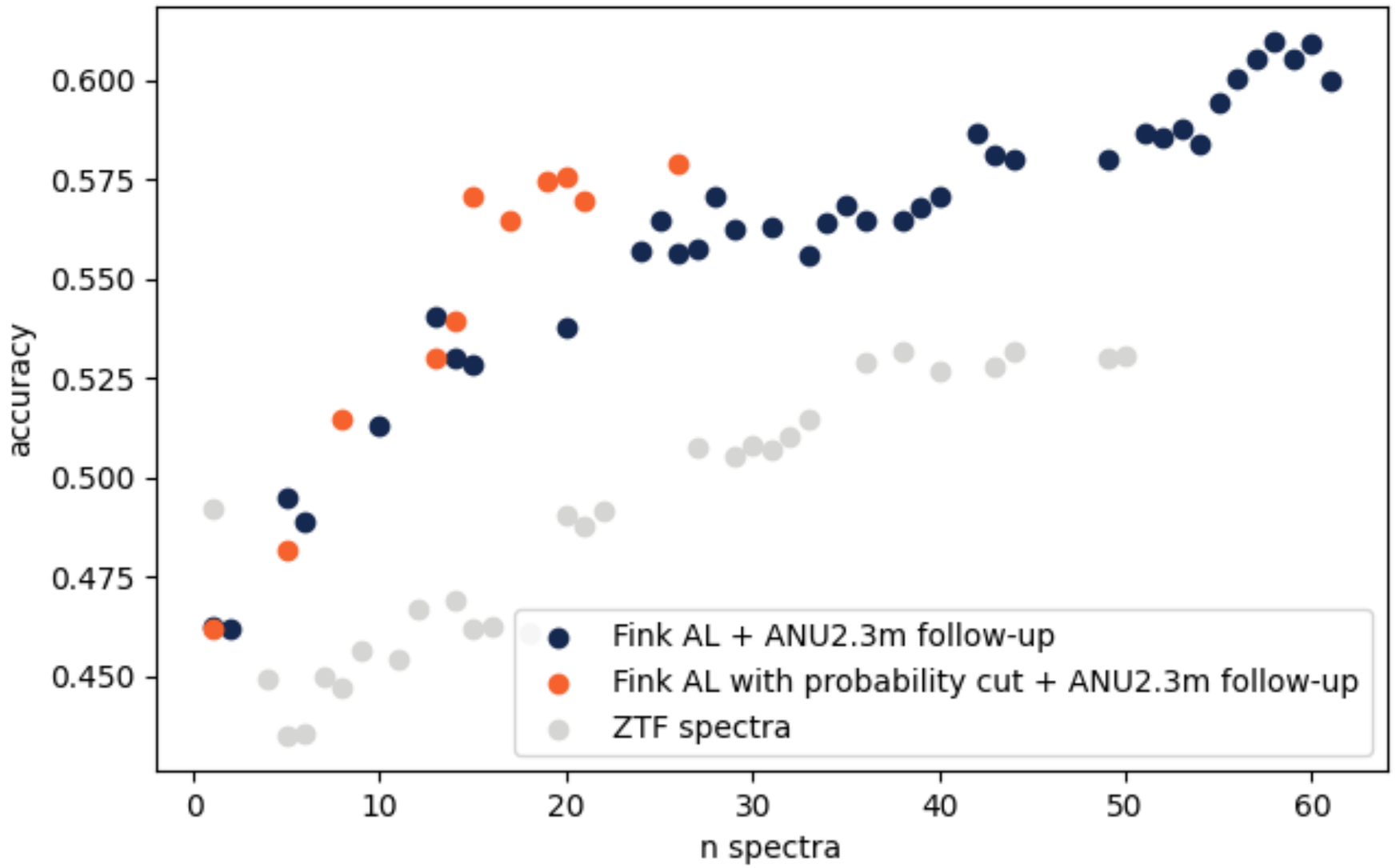
Preliminary!!!



40 light-curves as initial training set

Möller et al. in prep.

Preliminary!!!



61 labels
26 labels
50 labels

Möller et al. in prep.

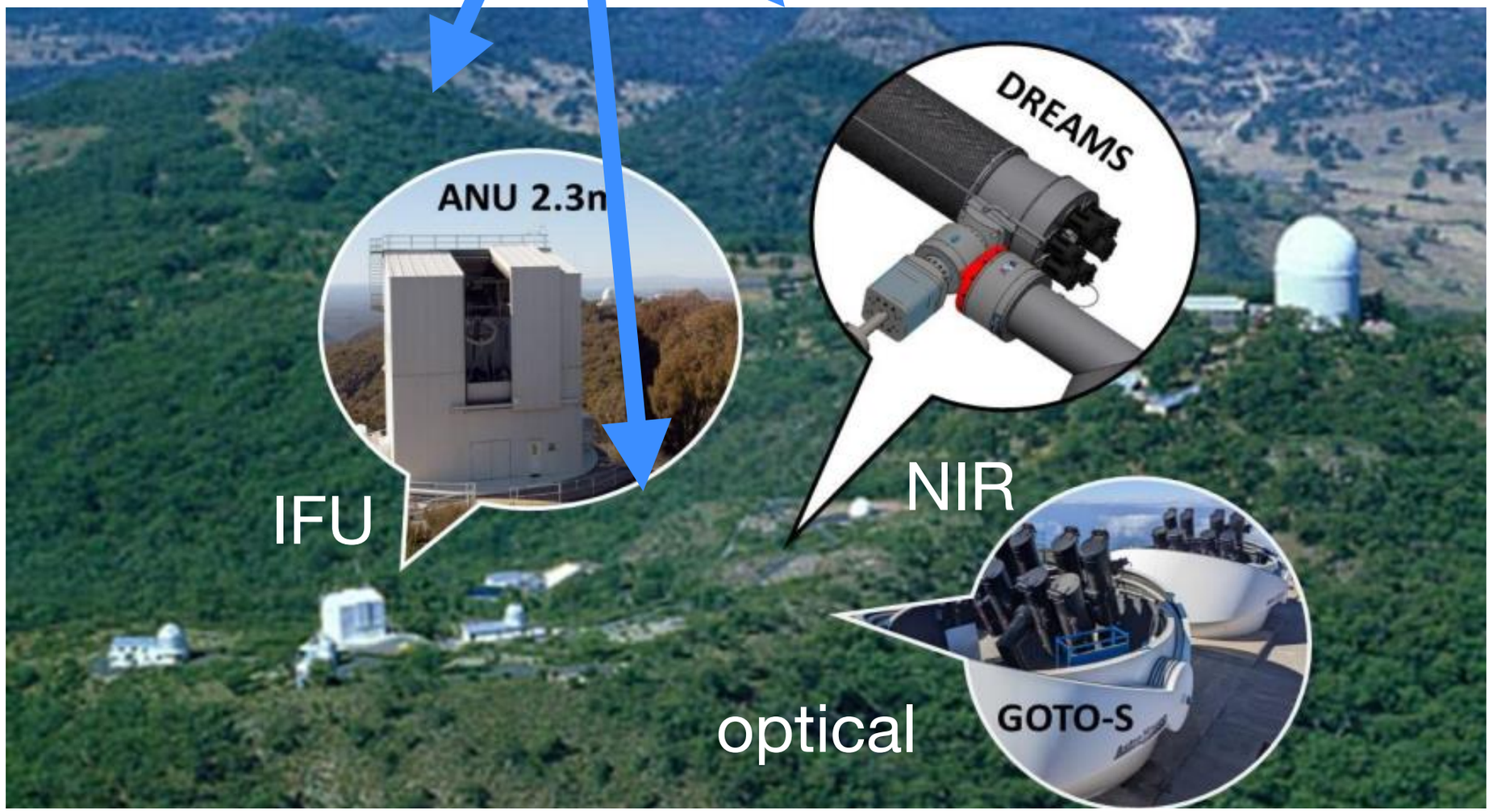
Robotic Network @ Siding Spring Observatory

Lidman, Colless, Wolf, Travouillon, Brough, Seitenzahl, Ruiter, Brown, Heger, Galloway, Möller, Kamath, O'Toole, Wen, Einecke

FINK

Alert System

TOM TOOLKIT Las Cumbres Observatory LCO



Automatic data reduction

Rubin is an amazing opportunity and a big data challenge!

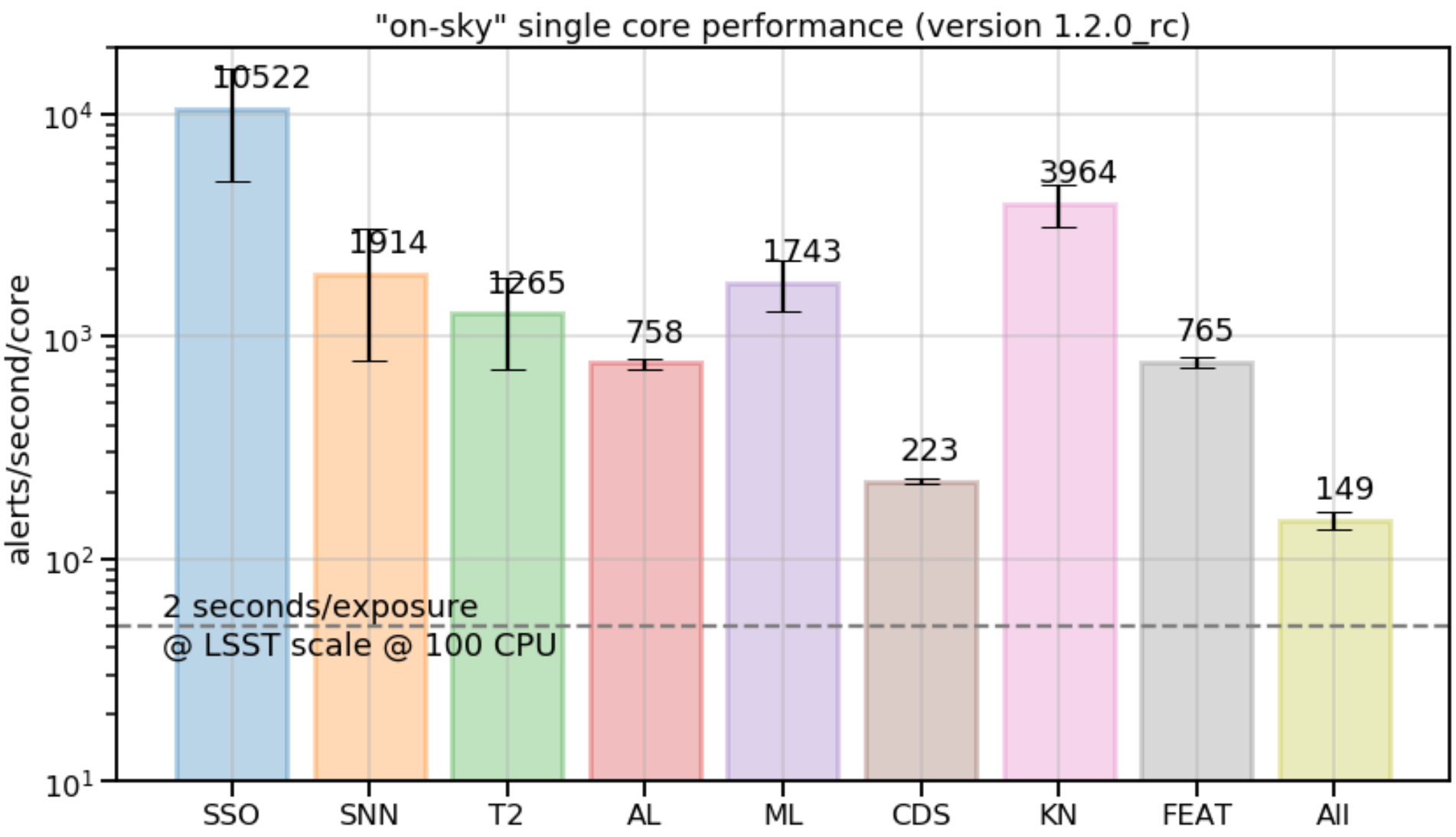
- All our data is public fink-portal.org
- First candidate ID modules: SNe, KNe, SSO, GRB, microlensing…

Supernova ML challenges:

- AL is a good strategy to improve training sets (optimise follow-up)
- This will percolate to simulations (built with templates)
- And then we can use large NNs accurately for precision science!

Community driven, join us!

Backup


"on-sky" single core performance (version 1.2.0_rc)
alerts/second/core
10^4
10^3
10^2
10^1
10522
1914
1265
758
1743
223
3964
765
149
2 seconds/exposure
@ LSST scale @ 100 CPU
SSO
SNN
T2
AL
ML
CDS
KN
FEAT
All